日　本　国　特　許　庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００２年１０月１６日

出願番号
Application Number: 特願２００２－３０１５３９

[ST.10/C]: [ＪＰ２００２－３０１５３９]

出願人
Applicant(s): インターナショナル・ビジネス・マシーンズ・コーポレーション

２００３年　４月１１日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号　出証特２００３－３０２５４７２

【書類名】　特許願
【整理番号】　JP9020132
【提出日】　平成14年10月16日
【あて先】　特許庁長官　殿
【国際特許分類】　G06F 15/401
【発明者】
【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内
【氏名】　吉田　一星
【特許出願人】
【識別番号】　390009531
【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレーション
【代理人】
【識別番号】　100086243
【弁理士】
【氏名又は名称】　坂口　博
【代理人】
【識別番号】　100091568
【弁理士】
【氏名又は名称】　市位　嘉宏
【代理人】
【識別番号】　100108501
【弁理士】
【氏名又は名称】　上野　剛史
【復代理人】
【識別番号】　100104880
【弁理士】
【氏名又は名称】　古部　次郎

【手数料の表示】

【予納台帳番号】　081504

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9706050

【包括委任状番号】　9704733

【包括委任状番号】　0207860

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　文書自動分類システム、不要語判定方法、文書自動分類方法、およびプログラム

【特許請求の範囲】

【請求項１】　学習用文書集合から語を抽出し、カテゴリごとに語のリストを作成するリスト作成手段と、

前記リスト作成手段により作成された前記リストを用いて、所定の語における各カテゴリでの出現頻度をもとにカテゴリごとの不要語を相対的に決定する不要語決定手段と

を含む文書自動分類システム。

【請求項２】　前記リスト作成手段は、記憶手段にある前記学習用文書集合からカテゴリごとに所定の語における出現頻度を示すリストを生成することを特徴とする請求項１記載の文書自動分類システム。

【請求項３】　前記不要語決定手段は、所定のカテゴリに属する語を取り出し、当該語が他のカテゴリにて所定の基準より多く出現する場合に不要語と決定することを特徴とする請求項１記載の文書自動分類システム。

【請求項４】　前記不要語決定手段は、前記所定のカテゴリから取り出される前記語が、予め定められる閾値および前記他のカテゴリに属する文書の個数によって決定される前記所定の基準より当該他のカテゴリにて多く出現する場合に不要語と決定することを特徴とする請求項３記載の文書自動分類システム。

【請求項５】　前記不要語決定手段により決定され、当該決定により不要語が除去された、カテゴリごとのリストを分類用カタログとして格納する分類用カタログ格納手段と、

前記分類用カタログ格納手段に格納された前記分類用カタログを用いて、分類対象文書に対して分類処理を施す文書分類手段と

を更に含む請求項１記載の文書自動分類システム。

【請求項６】　分野ごとに分類済みの文書を格納する分類済み文書集合格納装置と、

前記分類済み文書集合格納装置から取得された文書に含まれる単語の出現頻度

の情報を含む分野別のテーブルを作成する分野テーブル作成部と、

前記分野テーブル作成部により作成された前記分野別のテーブルから得られる所定の語における各分野での出現頻度に基づいて、当該テーブルから当該分野別に不要語を除去する不要語除去部と、

前記不要語除去部により不要語が除去された前記テーブルを格納する分類用カタログ格納装置と

を含む文書自動分類システム。

【請求項７】　分類される分類対象文書を格納する分類対象文書格納装置と、

前記分類対象文書格納装置に格納された前記分類対象文書に対し、前記分類用カタログ格納装置に格納された前記テーブルを用いて分類処理を行なう文書分類処理部と

を更に含む請求項６記載の文書自動分類システム。

【請求項８】　前記不要語除去部は、所定の分野に属する語を取り出し、当該語が他の分野にて所定の基準を超えて出現する場合に、当該語を不要語として前記テーブルから除去することを特徴とする請求項６記載の文書自動分類システム。

【請求項９】　前記分野テーブル作成部により作成される前記分野別のテーブルは、前記単語、当該単語の出現頻度、および当該単語の品詞に関する情報を含むことを特徴とする請求項６記載の文書自動分類システム。

【請求項１０】　文書自動分類システムにおける不要語判定方法であって、

学習用文書集合が格納されている記憶装置から文書に含まれる単語をカテゴリごとに抽出するステップと、

抽出された前記単語の出現頻度の情報を含むリストをカテゴリごとに作成するステップと、

作成された前記リストを用いて、所定のカテゴリに属する所定の単語における他のカテゴリでの出現頻度を認識するステップと、

認識された前記出現頻度に基づいて、カテゴリごとに不要語を判定するステップと

を含む不要語判定方法。

【請求項１１】　前記不要語を判定するステップでは、前記所定のカテゴリから選定された１つの単語に対して、前記他のカテゴリにて当該単語が所定の基準を超えて含まれているか否かによって不要語を判定することを特徴とする請求項１０記載の不要語判定方法。

【請求項１２】　前記所定の基準は、前記他のカテゴリ内の文書数および予め定められた所定の閾値により得られる値であることを特徴とする請求項１１記載の不要語判定方法。

【請求項１３】　前記所定の基準は、前記他のカテゴリ内での前記単語の頻度と、当該他のカテゴリ内での全ての単語の頻度合計とによって決定されることを特徴とする請求項１１記載の不要語判定方法。

【請求項１４】　文書自動分類システムにおける不要語判定方法であって、

記憶装置に格納されている分野別に分類済みの文書集合から分野別の単語に関する情報を取得し、

取得された前記単語に関する情報に基づいて、特定の分野に属する単語が他の分野に出現する頻度を認識し、

認識される前記頻度に基づいて前記単語が前記特定の分野を識別するのに不要な単語か否かを判定することを特徴とする文書自動分類方法。

【請求項１５】　前記不要な単語であると判定された単語を除去して文書の分類用カタログを生成し、

生成された前記分類用カタログを記憶装置に格納することを特徴とする請求項１４記載の文書自動分類方法。

【請求項１６】　前記記憶装置に格納された前記分類用カタログを用いて、分類対象文書に対して分類処理を施すことを特徴とする請求項１５記載の文書自動分類方法。

【請求項１７】　コンピュータに、

学習用文書集合が格納されている記憶装置から文書に含まれる単語をカテゴリごとに抽出する機能と、

抽出された前記単語の出現頻度の情報を含むリストをカテゴリごとに作成する

機能と、

作成された前記リストを用いて、所定のカテゴリに属する所定の単語における他のカテゴリでの出現頻度を認識する機能と、

認識された前記出現頻度に基づいて、カテゴリごとに不要語を判定する機能とを実現させるプログラム。

【請求項１８】　前記コンピュータに、

判定された前記不要語を用いて分類用のリストを生成する機能を更に実現させる請求項１７記載のプログラム。

【請求項１９】　コンピュータに、

記憶装置に格納されている分野別に分類済みの文書集合から分野別の単語に関する情報を取得する機能と、

取得された前記単語に関する情報に基づいて、特定の分野に属する単語が他の分野に出現する頻度を認識する機能と、

認識される前記頻度に基づいて前記単語が前記特定の分野を識別するのに不要な単語か否かを判定する機能と

を実現させるプログラム。

【請求項２０】　前記コンピュータに、

前記不要な単語であると判定された単語を除去して文書の分類用カタログを生成する機能と、

生成された前記分類用カタログを用いて、分類対象文書を分類する機能と

を更に実現させる請求項１９記載のプログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、文書データを自動的に分類するための文書自動分類システム等に係り、より詳しくは、不要語を効果的に取り除く文書自動分類システム等に関する。

【０００２】

【従来の技術】

近年、電子化された文書データ(テキスト)が大量に流通するのに伴い、例えば文書格納データベースに存在する大量の文書を自動分類する文書自動分類システムが注目されている。この文書自動分類システムは、学習機能と分類機能との２つの要素から構成される。これらの機能を実現するために、決定木、Neural Network、ベクトル空間モデルなど、様々なモデルが提案されている。何れの方法においても、各カテゴリや文書を特定付ける語を文書から抽出することが重要である。しかしながら、文書から単語を頻度順に取り出す場合には、カテゴリを一意に決定するために、有用でない語(不要語)が上位を占めてしまう。この不要語を学習・分類前に除去しておくことで、文書自動分類システムの分類性能を大きく改善することができる。

【０００３】

不要語には、大きく分けて、機能語と一般語の２種類が存在する。機能語は、語と語の関係を表す助詞、助動詞などを指す。この機能語は、カテゴリに存在しないものが多いので、語の品詞を調べたり、予め不要語リストを作成しておくことにより除去することができる。一方、一般語は、機能語以外に一般的に用いられる語を指す。この一般語は、機能語と異なり、語の頻度によって決定される場合が多く、与えられた文書集合中の出現頻度がある上限または下限を越えた語を不要語とする手法が一般的に用いられる。この上限、下限を決める手法として、語の出現頻度に関する経験法則をもとに、多過ぎる或いは少な過ぎる語を判定して除去するＺｉｐｆの法則などが知られている。

【０００４】

文書自動分類技術に関する従来技術として、例えば、分類済み文書から複数分野語を学習し、この複数分野語に注目して関連度テーブルや分類対象文書の単語の頻度情報を詳細化することで、分類対象文書の各分野への関連度をより詳細に分析し、類似する分野での分類精度を向上させるものが存在する(例えば、特許文献１参照)。また、不要単語を登録した不要語辞書を設け、新規単語に不要語辞書中の不要単語と同一のものが含まれているときに新規単語を削除し、不要単語が削除された新規単語に対して単語重要度を決定する技術について開示されている(例えば、特許文献２参照)。更に、精度の高い類似文書検索を行なうために

、出現頻度をカウントして不要語リストを自動的に作成し、一定の割合以上(以下)出現する語を削除することで、類似度算出精度の向上を図った技術が示されている(例えば、特許文献３参照)。

【０００５】

【特許文献１】

特開平１０－２５４８８３号公報(第４－５頁、１５頁、図１)

【特許文献２】

特開平１１－１２０１８３号公報(第３－４頁、図１)

【特許文献３】

特開平１１－２５９５１５号公報(第３－５頁、図３)

【０００６】

【発明が解決しようとする課題】

このように、精度の高い文書自動分類を実行するためには、文書中に存在する抽出すべき単語対象から不要語を排除することが好ましい。しかしながら、特許文献１では、まず、不要語除去という概念が存在せず、どの単語も最低１つは関連の強いカテゴリが存在することを前提としており、品詞の限定や不要語リストの作成を行なわない限り不要語がそのままリスト登録されてしまい、精度の高い分類が困難となる。また、関連テーブルを作成した後、改めて詳細関連度テーブルを作成しており、多くの記憶容量を必要としてしまう。

【０００７】

また、特許文献２では、予め用意された不要語リストとの照合による不要語除去がなされているが、対象となるカテゴリの集合ごとに不要語リストを作り直す必要があり、また、時代と共に変化する用語に対して十分に対処することができない。更に、特許文献３では、学習用文書全体における各語の出現頻度をカウントしているものの、頻度の基準値を設定してそれを超えた語を除去する方法に留まり、除去されない不要語が数多く残る可能性が高く、その一方で、不要語の判定を広く行なうと、分類のために有用な語まで除去されてしまうという問題があった。また、上述したＺｉｐｆの法則では、上限・下限を越えない語の中にも不要語が含まれていたり、逆に、上限・下限を超えた語の中にカテゴリを特徴付け

る重要な語が含まれている場合がある。

【０００８】

本発明は、以上のような技術的課題を解決するためになされたものであって、その目的とするところは、文書自動分類において、不要語を効果的に取り除くことにある。

【０００９】

【課題を解決するための手段】

かかる目的のもと、本発明は、文書を自動的にカテゴリに分類する文書自動分類システムにおいて、学習用文書集合から語を抽出し、カテゴリごとに語のリストを作成するリスト作成手段と、このリスト作成手段により作成されたリストを用いて、所定の語における各カテゴリでの出現頻度をもとにカテゴリごとの不要語を相対的に決定する不要語決定手段と、この不要語決定手段により決定され、不要語が除去された、カテゴリごとのリストを分類用カタログとして格納する分類用カタログ格納手段と、この分類用カタログ格納手段に格納された分類用カタログを用いて、分類対象文書に対して分類処理を施す文書分類手段とを含む。

【００１０】

ここで、このリスト作成手段は、記憶手段にある学習用文書集合からカテゴリごとに所定の語における出現頻度を示すリストを生成することを特徴とすることができる。また、この不要語決定手段は、所定のカテゴリに属する語を取り出し、語が他のカテゴリにて所定の基準より多く出現する場合に不要語と決定すれば、カテゴリ間の相対的な出現頻度をもとに不要語を判断することができ、効果的に不要語を除去することができる。更に、この不要語決定手段は、所定のカテゴリから取り出される語が、予め定められる閾値および他のカテゴリに属する文書の個数によって決定される所定の基準より他のカテゴリにて多く出現する場合に不要語と決定することを特徴とすることができる。

【００１１】

他の観点から把えると、本発明が適用される文書自動分類システムは、分野ごとに分類済みの文書を格納する分類済み文書集合格納装置と、この分類済み文書集合格納装置から取得された文書に含まれる単語の出現頻度の情報を含む分野別

のテーブルを作成する分野テーブル作成部と、この分野テーブル作成部により作成された分野別のテーブルから得られる所定の語における各分野での出現頻度に基づいて、テーブルから分野別に不要語を除去する不要語除去部と、この不要語除去部により不要語が除去されたテーブルを格納する分類用カタログ格納装置と、分類される分類対象文書を格納する分類対象文書格納装置と、この分類対象文書格納装置に格納された分類対象文書に対し、分類用カタログ格納装置に格納されたテーブルを用いて分類処理を行なう文書分類処理部とを含む。

【００１２】

一方、本発明は、文書自動分類システムにおける不要語判定方法であって、分野テーブル作成手段により、学習用文書集合が格納されている記憶装置から文書に含まれる単語をカテゴリごとに抽出し、抽出された単語の出現頻度の情報を含むリストをカテゴリごとに作成するステップと、不要語判定手段により、作成されたリストを用いて、所定のカテゴリに属する所定の単語における他のカテゴリでの出現頻度を認識し、認識された出現頻度に基づいて、カテゴリごとに不要語を判定するステップとを含む。

【００１３】

ここで、この不要語を判定するステップでは、所定のカテゴリから選定された１つの単語に対して、他のカテゴリにてこの単語が所定の基準を超えて含まれているか否かによって不要語を判定することを特徴とすれば、カテゴリの特徴付けに役立たない語を有効に取り除くことができる点で好ましい。また、この所定の基準は、他のカテゴリ内の文書数および予め定められた所定の閾値により得られる値であることを特徴とすることができる。また他の態様として、この所定の基準は、他のカテゴリ内での単語の頻度と、この他のカテゴリ内での全ての単語の頻度合計とによって決定されることを特徴とすることもできる。

【００１４】

更に他の観点から把えると、本発明が適用される文書自動分類方法は、記憶装置に格納されている分野別に分類済みの文書集合から分野別の単語に関する情報を取得し、取得された単語に関する情報に基づいて、特定の分野に属する単語が他の分野に出現する頻度を認識し、認識される頻度に基づいてこの単語が特定の

分野を識別するのに不要な単語か否かを判定し、不要な単語であると判定された単語を除去して文書の分類用カタログを生成し、生成された分類用カタログを記憶装置に格納し、この記憶装置に格納された分類用カタログを用いて、分類対象文書に対して分類処理を施すことを特徴とすることができる。

【００１５】

尚、これらの発明は、コンピュータに各機能を実現させるプログラムとして機能させることができる。より具体的には、コンピュータに、学習用文書集合が格納されている記憶装置から文書に含まれる単語をカテゴリごとに抽出する機能と、抽出された単語の出現頻度の情報を含むリストをカテゴリごとに作成する機能と、作成されたリストを用いて、所定のカテゴリに属する所定の単語における他のカテゴリでの出現頻度を認識する機能と、認識された出現頻度に基づいて、カテゴリごとに不要語を判定する機能と、判定された不要語を用いて分類用のリストを生成する機能を実現させるプログラムとして把握することができる。

【００１６】

また、本発明は、コンピュータに、記憶装置に格納されている分野別に分類済みの文書集合から分野別の単語に関する情報を取得する機能と、取得された単語に関する情報に基づいて、特定の分野に属する単語が他の分野に出現する頻度を認識する機能と、認識される頻度に基づいて単語が特定の分野を識別するのに不要な単語か否かを判定する機能と、不要な単語であると判定された単語を除去して文書の分類用カタログを生成する機能と、生成された分類用カタログを用いて、分類対象文書を分類する機能とを実現させるプログラムとして把握することができる。

【００１７】

尚、これらのプログラムとしては、コンピュータを顧客に対して提供する際に、コンピュータ装置の中にインストールされた状態にて提供される場合の他、コンピュータに実行させるプログラムをコンピュータが読取可能に記憶した記憶媒体にて提供する形態が考えられる。この記憶媒体としては、例えばＣＤ－ＲＯＭ媒体等が該当し、ＣＤ－ＲＯＭ読取装置等によってプログラムが読み取られ、フラッシュＲＯＭ等にこのプログラムが格納されて実行される。また、これらのプ

ログラムは、例えば、プログラム伝送装置によってネットワークを介して提供される形態がある。このプログラム伝送装置としては、例えば、ネット上のサーバに設けられ、プログラムを格納するメモリと、ネットワークを介してプログラムを提供するプログラム伝送手段とを備えている。

【００１８】

【発明の実施の形態】

以下、添付図面を参照し、本発明が適用される実施の形態について詳細に説明する。

図１は、本実施の形態が適用される文書自動分類システム１０の構成を示したブロック図である。この文書自動分類システム１０は、パーソナルコンピュータ(ＰＣ)等のコンピュータ装置によって展開され、ＨＤＤ(ハードディスクドライブ)などの外部記憶装置にて構成されて各種データを格納するデータ格納装置２０と、外部メモリから読み出されたアプリケーションプログラムによってＣＰＵにより実行される処理部３０とを備えている。実際には、処理部３０の各ブロック構成要素は、ＣＰＵの実行プログラムの読み込み領域として或いは実行プログラムの処理データを書き込む作業領域として利用される複数個のＤＲＡＭチップ等からなる内部記憶装置にて展開される。

【００１９】

データ格納装置２０は、分野(カテゴリ)の学習処理に用いられる分類済みの文書を格納する、学習用文書集合である分類済み文書集合格納装置２１、不要語が除去された後の分類用カタログが格納される分類用カタログ格納装置２２、実際に文書分類処理がなされる対象となるテキストを格納する分類対象文書格納装置２３、分類された結果を格納する分類結果格納装置２４を備えている。この分類結果格納装置２４の内容は、分類済み文書集合格納装置２１に格納されて学習処理に利用されるように構成することもできる。ここで、「不要語」とは、例えば、カテゴリ(分野)の特徴付けに役立たない語と定義できる。

【００２０】

処理部３０は、不要語削除前に選択された分野(カテゴリ)ごとに、語のリストであるテーブル情報を生成する分野テーブル作成部３１、分野テーブル作成部３

１にて作成された分野テーブルの単語について、不要語の決定と決定された不要語を除去する処理とを実行する不要語決定・除去部３２、実際に文書分類処理を実行する文書分類処理部３３を備えている。

【００２１】

分野テーブル作成部３１は、分類済み文書集合格納装置２１から得られた文書を用いて、例えば単語の出現頻度等の情報を含むテーブルを作成し、テーブル情報として内部記憶装置に登録している。分類済み文書集合格納装置２１では、学習用文書である複数の文書が、例えば、「政治」「経済」「スポーツ」等の分野(カテゴリ)の集合に分類された状態にて格納されている。分野テーブル作成部３１では、このカテゴリの集合に分類された文書を読み込み、その文書を解析し、例えば、文書に含まれる単語(語)の出現頻度をカウントして、分野テーブルを生成している。テーブルのデータ量が多い場合には、外部記憶装置であるデータ格納装置２０に別途、格納するように構成することができる。尚、分類済み文書集合格納装置２１の代わりに、所定のネットワークを介して、学習用文書集合(分類済み文書集合)を取得するように構成することも可能である。

【００２２】

不要語決定・除去部３２では、分野テーブル作成部３１にて作成された分野テーブルを用いて、カテゴリ間の相対的な出現頻度によって不要語を決定する処理が実行される。不要語決定・除去部３２によって不要語が除去された分野テーブルは、分類用カタログ格納装置２２に格納される。

【００２３】

文書分類処理部３３では、分類対象文書格納装置２３に格納されている実際の分類対象となる文書に対し、分類用カタログ格納装置２２に格納された分類用カタログ(不要語が除去された分野テーブル)を用いて、文書分類処理が実行される。この文書分類処理部３３による分類結果は、分類結果格納装置２４に格納される。

【００２４】

ここで、分野テーブル作成処理について説明する。

図２は、分野テーブル作成部３１にてなされる処理を示したフローチャートで

ある。分野テーブルの作成に際し、分野テーブル作成部３１では、分類済み文書集合格納装置２１に格納されている全ての分野について作業がなされているか否かが判断される(ステップ１０１)。全ての分野についての作業がなされていない場合には、まず、分野を１つ選び(ステップ１０２)、作業していない文書が分野内にあるか否かの判断がなされる(ステップ１０３)。分野内にない場合には、ステップ１０１に戻り、まだ残っている場合には、その分野から文書を１つ選ぶ(ステップ１０４)。そして、作業していない単語が文書内にあるか否かが判断され(ステップ１０５)、もう残っていない場合には、ステップ１０３へ戻り、未処理の単語がまだ文書内にある場合には、文書から単語が１つ選ばれる(ステップ１０６)。この単語の抽出では、形態素解析が用いられる。また、品詞によるフィルタリングをこのタイミングで行なうこともできる。

【００２５】

そして、単語が既にテーブル(分野テーブル)に登録されているか否かが判断され(ステップ１０７)、登録されている場合には、テーブル上の、登録単語の頻度(出現頻度)を１増やして、ステップ１０５に戻る。登録されていない単語である場合には、その単語をテーブルに登録し(ステップ１０９)、ステップ１０５へ戻る。このテーブル(分野テーブル)には、単語とその出現頻度の他に、各単語ごとに情報を持つこともできる。例えば、単語の品詞情報などを持つことができ、かかる場合には、この品詞情報等もテーブルに登録される。これらの一連の処理を行い、ステップ１０１にて全ての分野について作業をしたと判断された場合には、分野テーブル作成処理が終了する。

【００２６】

図３は、図２にて説明したような分野テーブル作成部３１にて作成されメモリに格納されるテーブル例を示した図である。ここでは、「スポーツ」の分野について、不要語除去前のテーブル例が示されている。テーブル情報は、単語を特定する番号である単語ＩＤごとに、単語、単語の品詞、単語の出現頻度が表されている。この単語の出現頻度は、「学習用の文書集合中に出現した回数の総計」を表している。１文書内に２個以上出現した場合も、その個数の分だけ数えている。尚、図３に示す例では、予め「名詞」と「動詞」だけをテーブルに登録すると

いう、前処理を行なってできたテーブルの模式図である。

【００２７】

次に、不要語除去処理について説明する。

図４は、不要語決定・除去部３２にてなされる処理を示したフローチャートである。不要語決定・除去部３２では、分野テーブル作成部３１にて作成された分野テーブルを用いて、全ての分野について作業がなされているか否かが判断される(ステップ２０１)。全ての分野についての作業がなされていない場合には、まず、１つの分野(Ａとする)を選ぶ(ステップ２０２)。そして、Ａの分野テーブルにおける全ての単語について作業したか否かが判断され(ステップ２０３)、全ての単語について作業した場合には、ステップ２０１に戻り、まだ残っている場合には、Ａの分野テーブルから１つの単語(Ｗとする)を選ぶ(ステップ２０４)。そして、Ａ以外の全ての分野と比較したか否かが判断され(ステップ２０５)、Ａ以外の全ての分野と比較した場合には、ステップ２０３へ戻り、比較していない場合には、Ａ以外の分野から１つの分野(Ｂとする)を選ぶ(ステップ２０６)。そして、予め定められている基準に対し、この基準を超えて、Ｂの分野テーブルにＷが含まれているか否かが判断され(ステップ２０７)、基準を超えて含まれていない場合には、ステップ２０５からの処理に戻る。含まれている場合には、このＷを不要語と判定して(ステップ２０８)、ステップ２０３からの処理に戻る。ステップ２０１にて、全ての分野について作業がなされたと判断された場合には、不要語除去の処理は終了し、除去結果のテーブル情報が分類用カタログ格納装置２２に格納される。

【００２８】

即ち、図４に示す不要語除去の方法では、所定のカテゴリＡに属する語(単語)Ｗを１つ、取り出し、この語Ｗが他のカテゴリＢの中で、所定の基準より多く出現するならば、この語Ｗがカテゴリの不要語と決定している。これを、カテゴリＡに属する全ての語について行なう。また、この一連の処理を、カテゴリＡ以外のカテゴリに対して、それぞれの判断対象となるカテゴリの役割を取り替えて、全てのカテゴリに対して不要語を決定している。

【００２９】

ここで、ステップ２０７の判断である「基準を超えて含まれている」を定義する方法としては、幾つかの方法が考えられる。例えば、後述するように、ある閾値を定め、分類済み文書集合格納装置２１に格納された学習用文書の個数に対し、

文書数×閾値

で得られた値に対してＢ内での単語Ｗの頻度が超える場合には、「基準値を超えて含まれる」と定義することができる。また、他の例として、例えば、

単語ＷのＢ内での頻度 ÷ Ｂ内の全ての単語の頻度合計

がある閾値を超えた場合には、「基準を超えて含まれる」と定義するように構成することもできる。

【００３０】

尚、図４に示す不要語除去の方法は、他の既存の不要語除去手法と組み合わせて用いることも可能である。また、分野(カテゴリ)が階層構造を成している場合も、同一階層に存在する分野に対してこのアルゴリズムを適用することによって、拡張することもできる。

【００３１】

図５(ａ)～(ｃ)は、この不要語処理のアルゴリズムについて更に詳しく説明するための図である。本アルゴリズムでは、まず、閾値Ｒ(０≦Ｒ≦１)が処理部３０に格納される。図５(ａ)～(ｃ)に示す例では、この閾値として、「０.０５」という値が記憶されている。また、図５(ａ)～(ｃ)に示す例では、分野(カテゴリ)として、スポーツ、経済、政治、の３分野について示され、それぞれの学習用文書の文書数は、８０文書、１００文書、１５０文書であるものとする。更に、図５(ａ)～(ｃ)に示す各カテゴリに属する語Ｗは、各カテゴリに属する文書の中に存在する語であり、その数値は、文書に含まれる語の頻度を示す。ここで、或る語の頻度として、例えば「その語のカテゴリに出現する個数の総計」や「カテゴリ内の、その語を含む文書の個数」など、任意の指標を採用することができる。

【００３２】

図５(ａ)に示すように、まず、カテゴリ「スポーツ」の中で頻度が５０である

単語「日本」を不要語とするか否かの判断を行なう。従来では、ただ単に、この頻度５０が大きいか小さいかだけを判断対象としていたが、本実施の形態では、他のカテゴリにおける頻度の状況を確認し、カテゴリ間の相対的な出現頻度をもとに不要語を決定している。そのために、他のカテゴリである「経済」の文書の中で、単語「日本」がどの程度、使用され、出現しているか、を判断している。より具体的には、カテゴリ「経済」の文書数を閾値Ｒで掛け合わせた値(１００×０．０５＝５)と、単語「日本」の頻度(３０個)とが比較される。３０は５よりも大きい(３０＞５)ことから、「スポーツ」で用いられる単語「日本」は、他のカテゴリ(例えば「経済」)でも頻繁に用いられる可能性がある単語と考えられる。従って、実際に文書の分類作業を行なう際、「スポーツ」のカテゴリを判断する判断対象として「日本」は好ましくないと考えられる。そこで、カテゴリ「スポーツ」では、単語「日本」が不要語とされる。

【００３３】

次に、図５(ｂ)に示すように、カテゴリ「スポーツ」の中で単語「代表」を不要語とするか否かの判断を行なう。まず、他のカテゴリの１つである「経済」の中で、単語「代表」の頻度は２であり、カテゴリ「経済」の文書数を閾値Ｒで掛け合わせた値(１００×０．０５＝５)と比較して小さい(２＜５)ことから、この段階では、カテゴリ「スポーツ」での不要語とは判断しない。しかしながら、もう一方の他のカテゴリである「政治」の中で、単語「代表」の頻度は８である。このとき、カテゴリ「政治」の文書数を閾値Ｒで掛け合わせた値(１５０×０．０５＝７．５)と比較して、出現頻度が大きい(８＞７．５)ことが理解できる。その結果、カテゴリ「スポーツ」の中の単語「代表」は、他のカテゴリの状態を判断して、識別単語としては好ましいものとは判断できない。そこで、「スポーツ」の中の単語「代表」は、不要語であると判断される。

【００３４】

更に、図５(ｃ)に示すように、カテゴリ「スポーツ」の中で単語「選手」を不要語とするか否かの判断を行なう。まず、他のカテゴリの１つであるカテゴリ「経済」の中で、単語「選手」の頻度は３であり、カテゴリ「経済」の文書数を閾値Ｒで掛け合わせた値(１００×０．０５＝５)と比較して小さい(３＜５)ことか

ら、単語「選手」は、カテゴリ「スポーツ」での不要語とは判断しない。また、もう一方の他のカテゴリであるカテゴリ「政治」の中で、単語「選手」の頻度は１である。カテゴリ「政治」の文書数を閾値Ｒで掛け合わせた値(１５０×０．０５＝７．５)と比較して小さい(１＜７．５)ことが理解できる。従って、カテゴリ「スポーツ」の中の単語「選手」は、他のカテゴリにて出現頻度が低く、識別単語として好ましいものと判断され、「スポーツ」の中の単語「選手」は不要語ではなく、削除されずに残される。

【００３５】

図６は、図５(ａ)～(ｃ)によって、全てのカテゴリに対して不要語を除去した後の状態を説明するための図である。上述したようなアルゴリズムによる不要語除去作業を全てのカテゴリに対して施す。図６において、斜線で示される領域に存在する単語は、不要語として除去される単語である。カテゴリ「スポーツ」では単語「日本」、「代表」が、カテゴリ「経済」では単語「日本」、「選手」、「代表」が、カテゴリ「政治」では単語「日本」、「代表」、「銀行」、「選手」が、不要語として除去される。

【００３６】

図７は、図３にて示した、分野テーブル作成部３１にて作成されメモリに格納されるテーブル例から、不要語を除去した後の分野テーブルの例を示した図である。図３と同様に、「スポーツ」の分野を例に挙げている。テーブル情報は、不要語除去後に残った単語を特定する番号である単語ＩＤごとに、単語、単語の品詞、単語の出現頻度が表されている。尚、図３と同様に、単語の出現頻度は、「学習用の文書集合中に出現した回数の総計」を表している。図７に示すような、不要語決定・除去部３２にて不要語が除去された分野テーブルは、分類用カタログ格納装置２２に分類用カタログとして格納される。尚、分類用カタログ格納装置２２に格納するに際し、図７に示すような、不要語が取り除かれた語のリストをそのまま用いてもよく、または、このリストに既存の「語の重み付け手法」を用いてリストを改良して格納することもできる。

【００３７】

以上のようにして不要語が除去された結果を用いて、実際に文書分類処理が実

行される。不要語を除去して得られた分野テーブルを文書分類処理に適用する方法としては幾つかのものが考えられるが、ここでは、「ベクトル空間モデル」と呼ばれる方法を例に挙げて説明する。

【００３８】

分類用カタログ格納装置２２には、不要語除去を経て作成された分野テーブルが格納されているが、分野(カテゴリ)ごとに、「語」と「語の重み」のペアが登録されている。図６に示す例では、「スポーツ」のカテゴリにて、語である「選手」、語の重みとして「２０」が登録されている。例えば、図６に示すような場合には、「選手」、「取引」、「銀行」、「ビール」、「首相」という５個の単語(語)の組を基底とするベクトル空間を考え、この空間内で「文書と各分野との距離」を計算する。尚、複数の分野に出てくる場合には、重複して出てきている語をまとめて１個としてベクトル空間が作成される。図６に示す例では、各分野のベクトルは、以下のようになる。

スポーツ　：（２０，０，０，０，０）

経済　　：（０，２０，１０，３，０）

政治　　：（０，０，０，０，１００）

【００３９】

次に、分類対象の文書から、文書ベクトルを作成する方法を説明する。ここでは、まず、分類対象文書格納装置２３から得られる分類対象の文書Ｄを形態素解析し、単語とその出現頻度との組をテーブルにする。例えば、

分類対象の文書の内容：

「Ａ国の首相が、Ｂ国の首相とイラク問題について会談した。」

について、形態素解析を行い、下記のようなテーブルを作成する。

（Ａ，１）、（国，２）、（首相，２）、（イラク，１）、（問題，１）、（会談，１）

次に、このようにして作成されたテーブルと、既に作成されているベクトル空間の基底とを比較し、ベクトル空間の基底になっている(登録されている)単語の情報のみを用いてベクトルを作成し、分類対象文書のベクトルが生成される。上記の例では、生成される文書ベクトルは、

選手、　取引、　銀行、　ビール、首相

（　０、　　０、　　０、　　０、　　２　）

となる。

【００４０】

その後、「文書と各分野との距離」の計算に、以上のようにして生成されたベクトルのなす角度の余弦が用いられる。

図８（ａ），（ｂ）は、本実施の形態にて利用されるベクトル空間モデルを説明するための図である。この余弦は、図８（ａ）に示すベクトルＡとベクトルＢに対して、その角度をθとすると、

$$\cos\theta = (A \cdot B) \div (|A||B|)$$

で定義される。ここで、Ａ・ＢはＡとＢとの積、｜Ａ｜はＡのノルム（長さ）を表す。余弦の値、即ちｃｏｓθは、０と１の間をとり、１に近いほどθが小さくなる。つまり、ｃｏｓθの値が大きいほど、ＡとＢとは互いに「近い」と考えられる。

【００４１】

文書の分類において、余弦は、次のようにして用いることができる。分類したい文書に対応するベクトルをＡ、分野に対応するベクトルをＢとし、各Ｂに対して、ＡとＢとの余弦を計算する。Ａに対して余弦の値を最も大きくするようなＢの分野を、Ａが属する分野と判定すればよい。図８（ｂ）に示すように、分類対象文書をベクトルＡにとり、政治、経済、スポーツの各カテゴリをベクトルＢにとる。そして、分類対象文書と政治、分類対象文書と経済、分類対象文書とスポーツ、の各々の余弦を、上述した式にて算出する。図８（ｂ）に示す例では、分類対象文書と政治との角度が最も小さく、余弦が最も大きくなり、分類対象文書は「政治」のカテゴリに属するものと判定することができる。

【００４２】

図９は、このようなベクトル空間モデルを用いて文書分類処理部３３にて実行される文書分類処理の流れを示したフローチャートである。文書分類処理部３３では、まず、分類対象文書格納装置２３から分類対象文書Ｄが取得される（ステップ３０１）。次に、分類対象文書Ｄの単語を全て抽出し、分類対象文書Ｄに対応するベクトルＶｄを作成する（ステップ３０２）。ここで、全ての分野について

作業したか否かが判断され(ステップ３０３)、作業が残っている場合には、分野を１つ選んでＡとする(ステップ３０４)。そして、ベクトルＶｄと、Ａに対応するベクトルＶａとの距離を、上述のようにして計算する(ステップ３０５)。ステップ３０３へ戻り、全ての作業について終了した場合には、計算した距離を用いて、分類対象文書Ｄの分類先の分野を決定し(ステップ３０６)、分類結果格納装置２４に結果を格納して処理が終了する。

【００４３】

以上、詳述したように、本実施の形態では、文書自動分類における不要語を「他のどれかのカテゴリにもある程度以上含まれている語」という定義を行い、カテゴリ間の相対的な出現頻度から不要語除去を行なっている。これによって、カテゴリの特徴付けに役立たない語(不要語)を新たに定義することができ、この定義によって、従来の手法に比べて、より効果的に不要語を除去することができる。また、不要語が除去されたリストを分類用カタログ格納装置２２に格納し、このリストを用いて実際の文書分類処理を実行することで、実際の文書処理に際して不要語か否かを判断するといった手間を省くことができる。即ち、実際の分類対象文書を解析して不要語を除去する必要がなく、分類作業を迅速化することが可能となる。

【００４４】

【発明の効果】

以上説明したように、本発明によれば、文書の自動分類において、不要語を効果的に取り除くことが可能となる。

【図面の簡単な説明】

【図１】　本実施の形態が適用される文書自動分類システムの構成を示したブロック図である。

【図２】　分野テーブル作成部にてなされる処理を示したフローチャートである。

【図３】　図２にて説明したような分野テーブル作成部にて作成されメモリに格納されるテーブル例を示した図である。

【図４】　不要語除去部にてなされる処理を示したフローチャートである。

【図５】　(ａ)～(ｃ)は、この不要語処理のアルゴリズムについて更に詳しく説明するための図である。

【図６】　図５(ａ)～(ｃ)によって、全てのカテゴリに対して不要語を除去した後の状態を説明するための図である。

【図７】　図３にて示した分野テーブル作成部にて作成されメモリに格納されるテーブル例から不要語を除去した後の分野テーブルの例を示した図である。

【図８】　(ａ),(ｂ)は、本実施の形態にて利用されるベクトル空間モデルを説明するための図である。

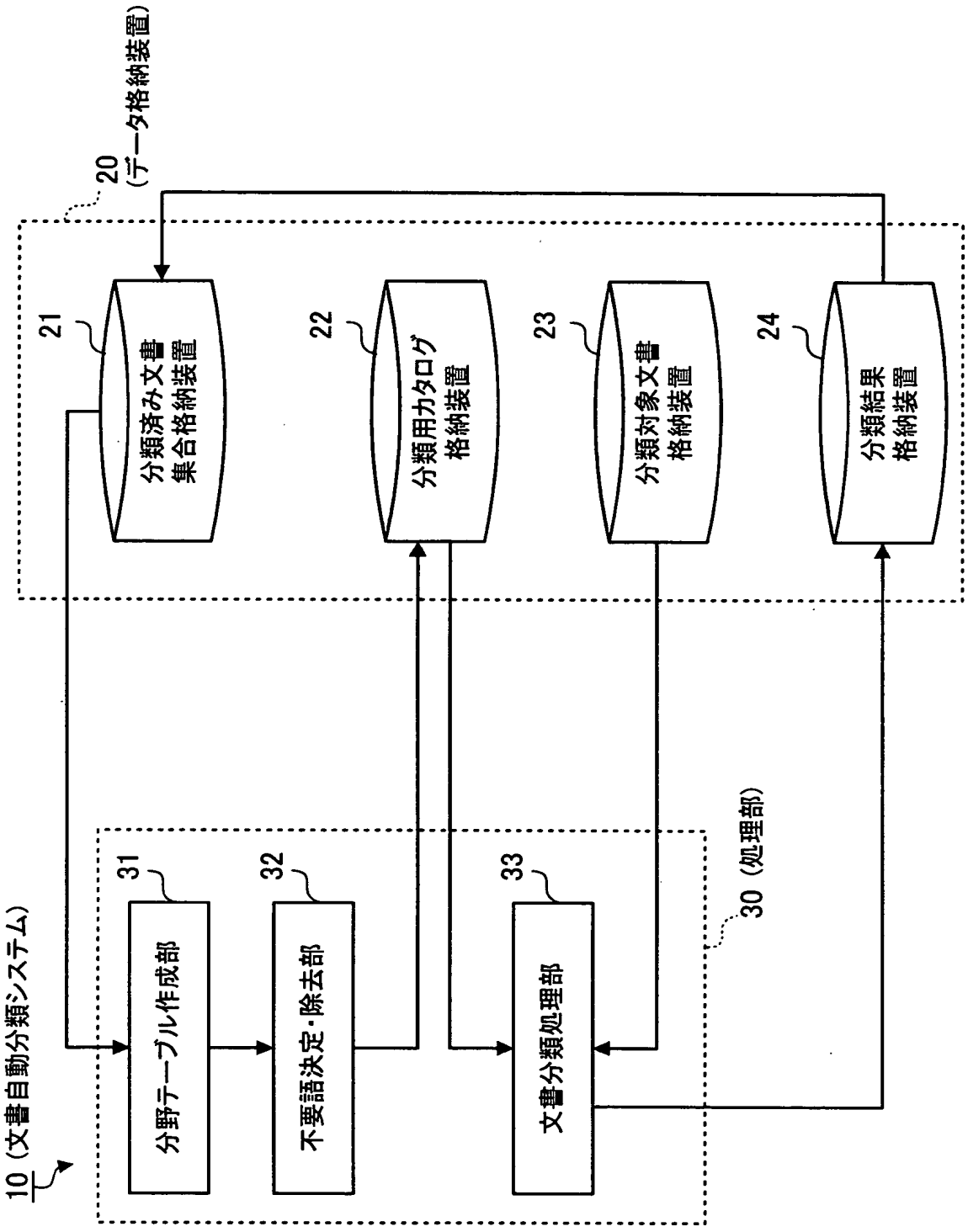
【図９】　ベクトル空間モデルを用いて文書分類処理部にて実行される文書分類処理の流れを示したフローチャートである。

【符号の説明】

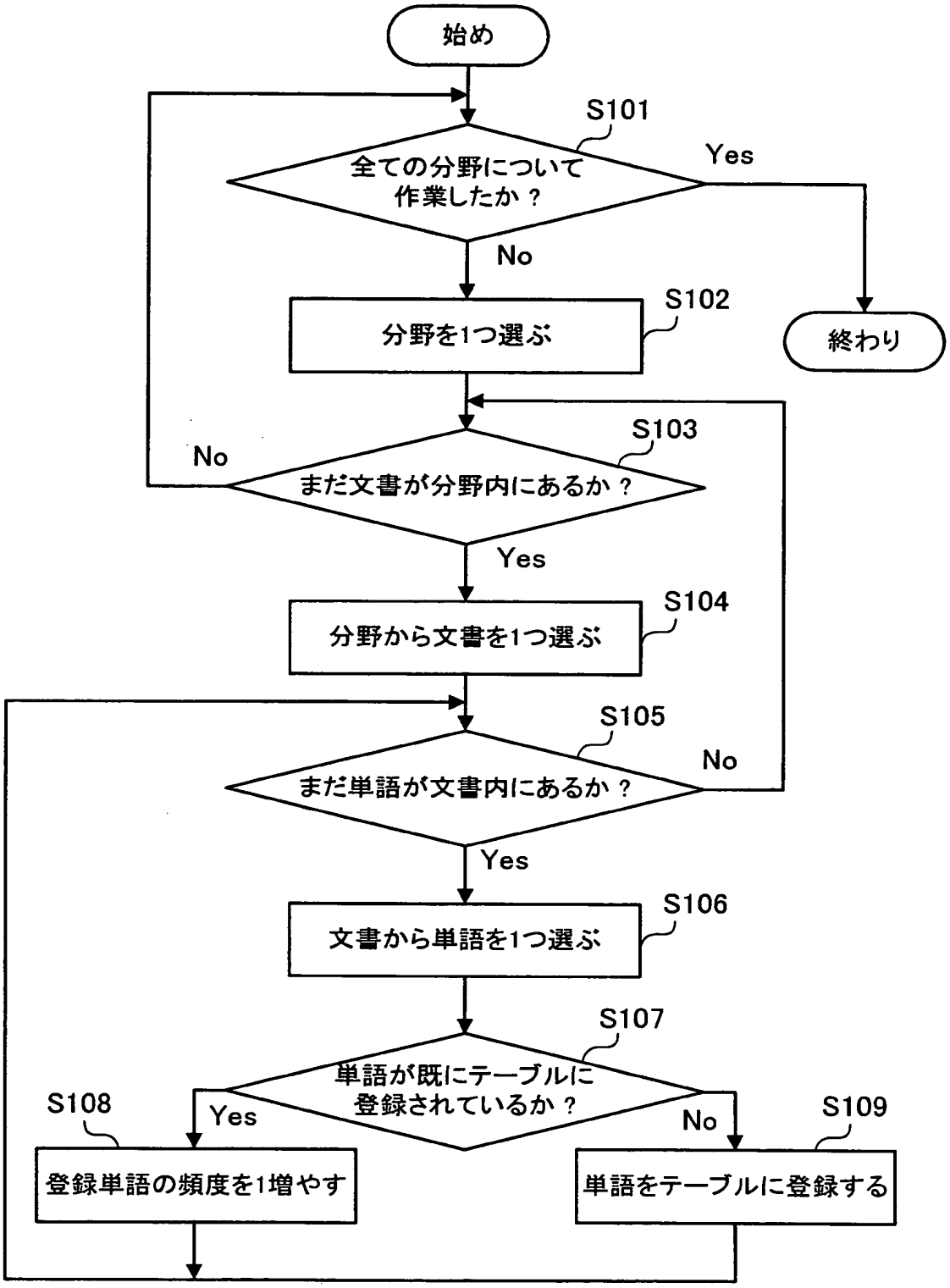
１０…文書自動分類システム、２０…データ格納装置、２１…分類済み文書集合格納装置、２２…分類用カタログ格納装置、２３…分類対象文書格納装置、２４…分類結果格納装置、３０…処理部、３１…分野テーブル作成部、３２…不要語決定・除去部、３３…文書分類処理部

【書類名】　　図面

【図１】


10 (文書自動分類システム)
20 (データ格納装置)
21
分類済み文書
集合格納装置
22
分類用カタログ
格納装置
23
分類対象文書
格納装置
24
分類結果
格納装置
30 (処理部)
31
分野テーブル作成部
32
不要語決定・除去部
33
文書分類処理部

【図２】


始め
S101
全ての分野について作業したか？
Yes
No
S102
分野を1つ選ぶ
終わり
S103
No
まだ文書が分野内にあるか？
Yes
S104
分野から文書を1つ選ぶ
S105
No
まだ単語が文書内にあるか？
Yes
S106
文書から単語を1つ選ぶ
S107
単語が既にテーブルに登録されているか？
S108
Yes
No
S109
登録単語の頻度を1増やす
単語をテーブルに登録する

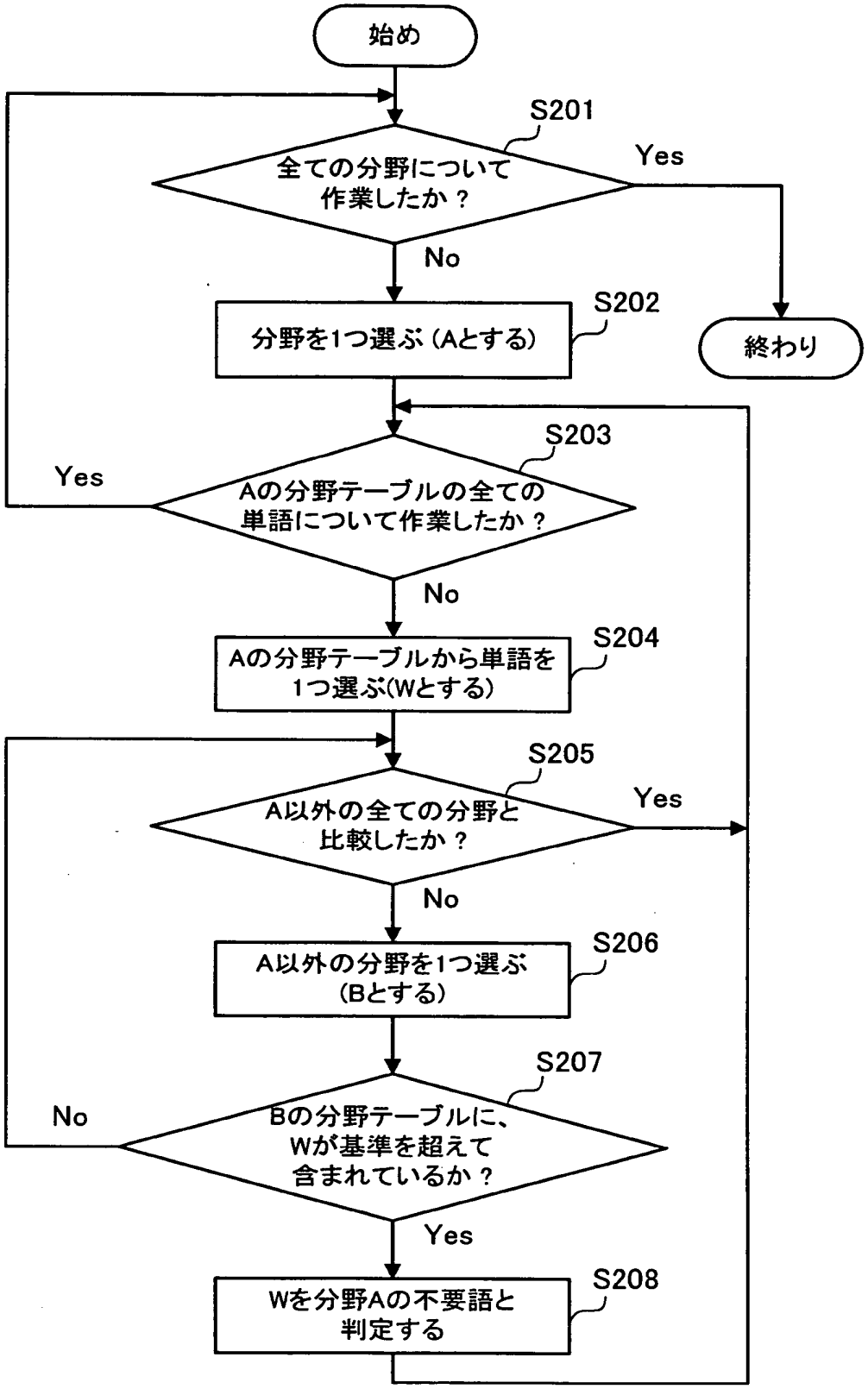
【図３】

分野テーブル(不要語除去前)

単語ID　単語　品詞　出現頻度

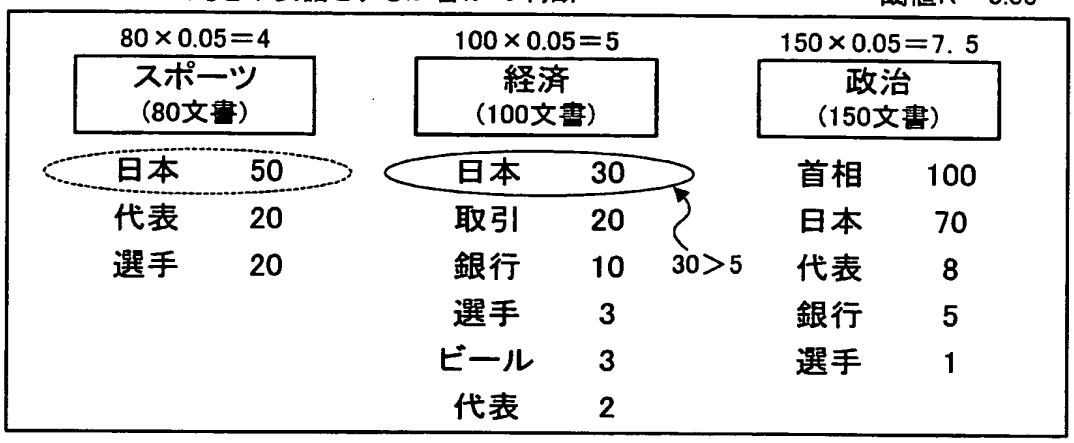
0：れ：動詞：8349
1：い：動詞：4841
2：な：動詞：4335
3：行：動詞：3570
4：日：名詞：2250
5：日本：固有名詞：1856
6：し：動詞：1716
7：チーム：名詞：1175
8：選手：名詞：1128
9：試合：名詞：714
10：い：動詞：660
11：大：名詞：583
12：通算：名詞：532
13：決勝：名詞：513
14：優勝：名詞：494
15：出場：動詞：456
16：リーグ：名詞：448
17：大会：名詞：432
18：打席：名詞：396
19：あ：動詞：357
20：連続：名詞：357
21：男子：名詞：345
22：決め：動詞：323
23：代表：名詞：312
24：打：動詞：306
25：今季：名詞：266
26：世界：名詞：252
27：話：動詞：238
28：敗れ：動詞：234
29：女子：名詞：221
30：投手：名詞：220
31：ついて：動詞：220
32：監督：名詞：216
33：外野手：名詞：216
34：○○○○：固有名詞：216
35：よ：動詞：208
36：安打：名詞：204
37：大リーグ：名詞：195
38：世界選手権：名詞：192
39：東京：固有名詞：190
40：選手権：名詞：182
41：出場：名詞：182
42：シドニー五輪：固有名詞：176
43：首位：名詞：168
45：せ：動詞：154
46：○○○：固有名詞：153
：
：
：
（以下、同様にして続く）

【図４】

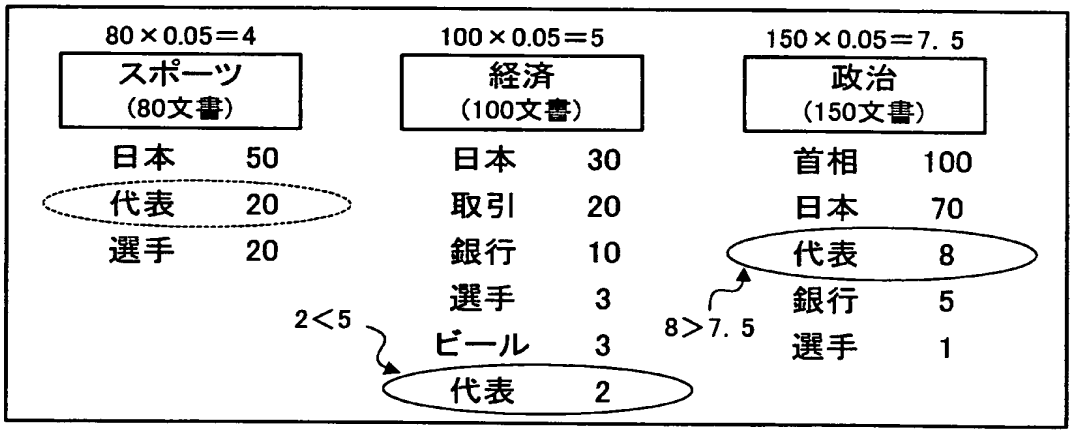

始め
S201
全ての分野について作業したか？
Yes
No
S202
分野を1つ選ぶ（Aとする）
終わり
S203
Yes
Aの分野テーブルの全ての単語について作業したか？
No
S204
Aの分野テーブルから単語を1つ選ぶ(Wとする)
S205
A以外の全ての分野と比較したか？
Yes
No
S206
A以外の分野を1つ選ぶ（Bとする）
S207
No
Bの分野テーブルに、Wが基準を超えて含まれているか？
Yes
S208
Wを分野Aの不要語と判定する

【図５】

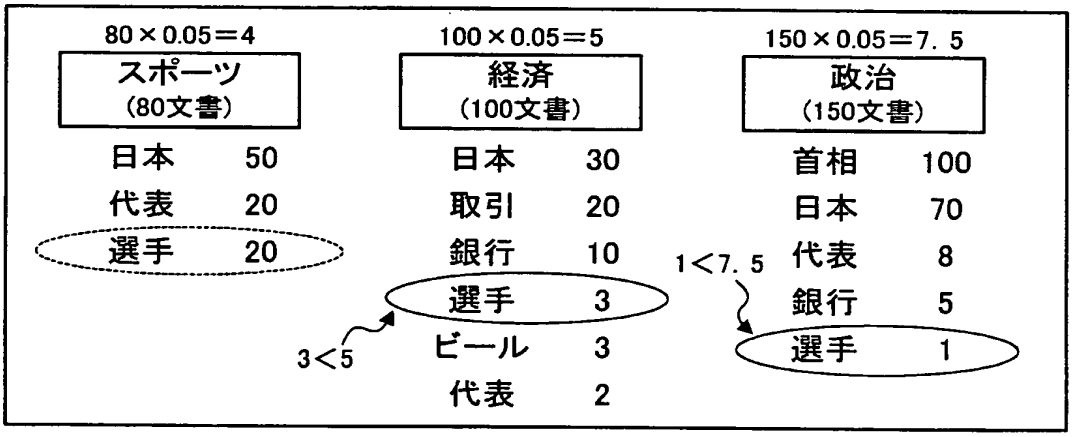
(a) スポーツの「日本」を不要語とするか否かの判断　　閾値R = 0.05



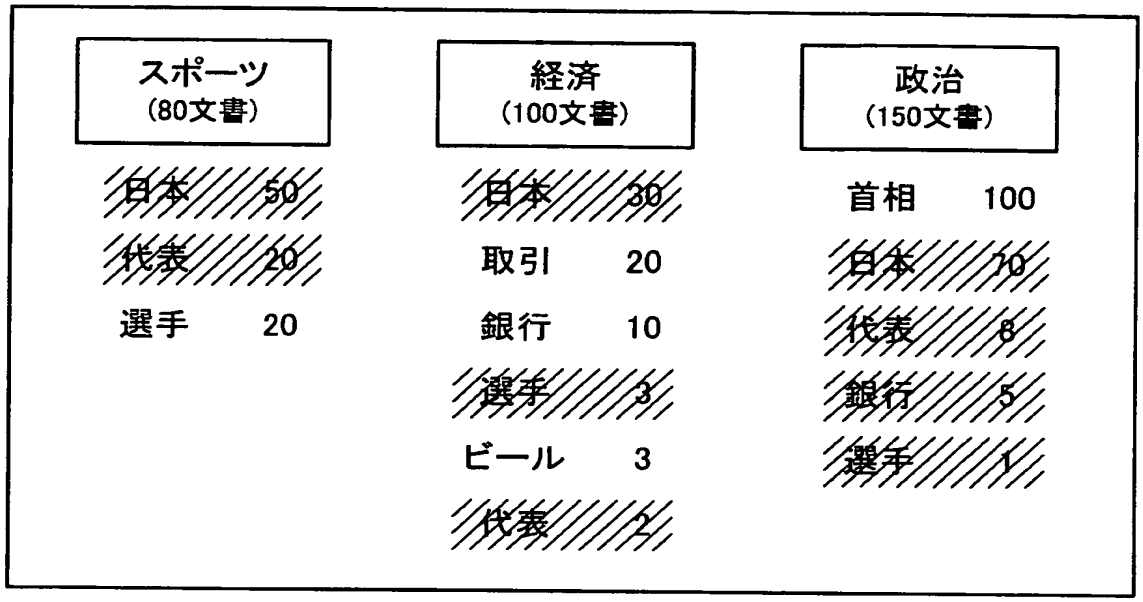
(b) スポーツの「代表」を不要語とするか否かの判断　　閾値R = 0.05



(c) スポーツの「選手」を不要語とするか否かの判断　　閾値R = 0.05

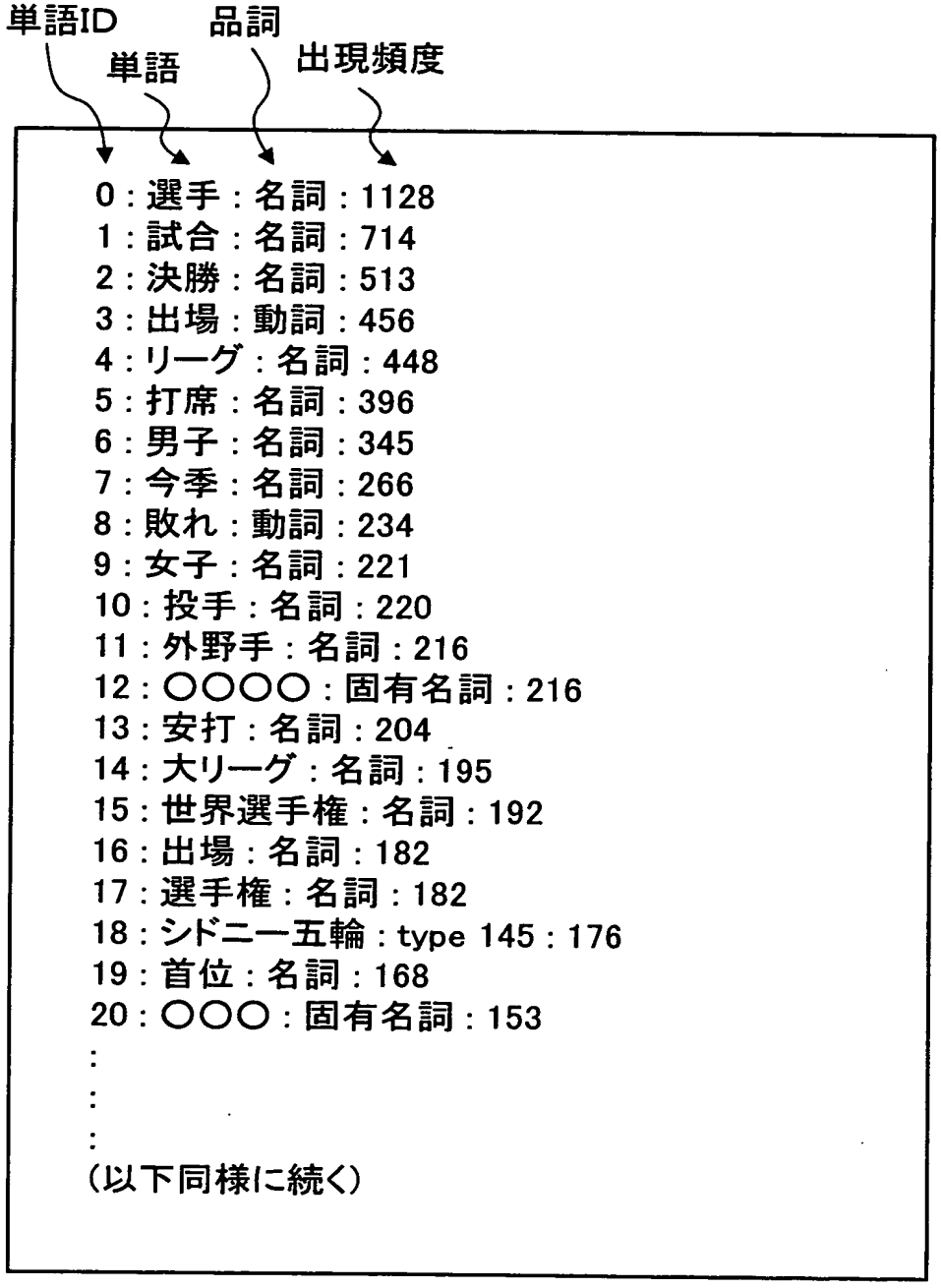


【図６】

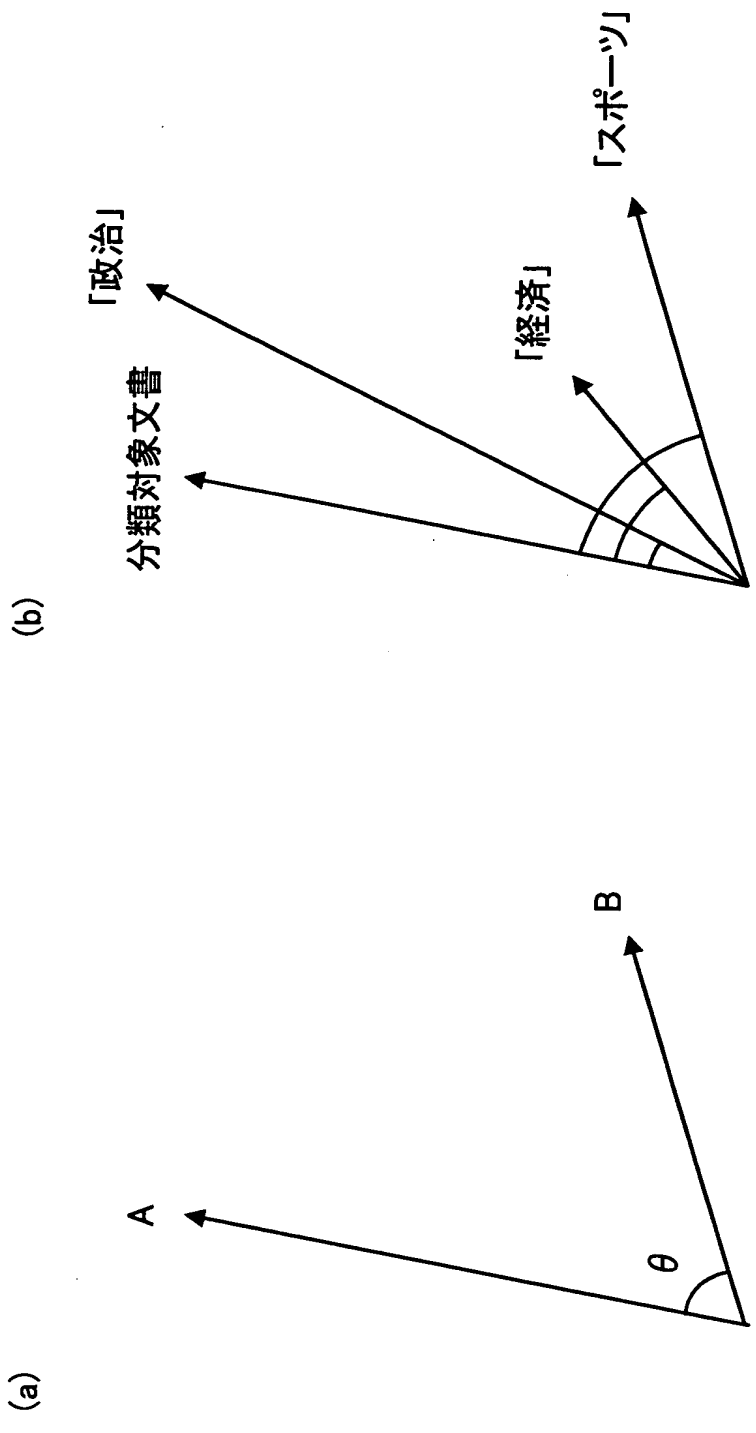

スポーツ
(80文書)
日本 50
代表 20
選手 20
経済
(100文書)
日本 30
取引 20
銀行 10
選手 3
ビール 3
代表 2
政治
(150文書)
首相 100
日本 70
代表 8
銀行 5
選手 1

【図７】

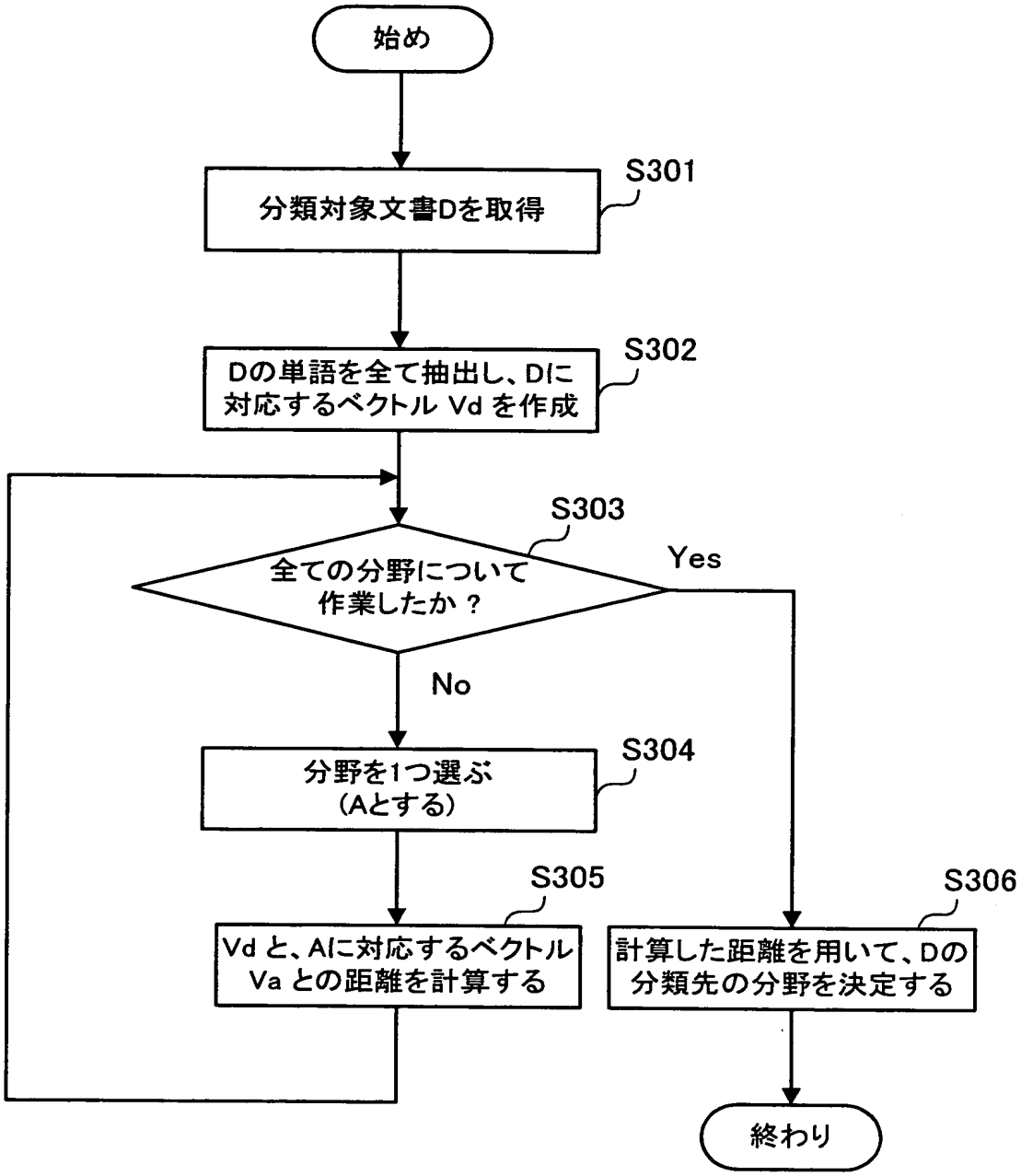
分野テーブル(不要語除去後)



【図８】



【図９】


始め
S301
分類対象文書Dを取得
S302
Dの単語を全て抽出し、Dに
対応するベクトル Vd を作成
S303
全ての分野について
作業したか？
Yes
No
S304
分野を1つ選ぶ
（Aとする）
S305
Vd と、Aに対応するベクトル
Va との距離を計算する
S306
計算した距離を用いて、Dの
分類先の分野を決定する
終わり

【書類名】　　要約書

【要約】

【課題】　文書自動分類において、不要語を効果的に取り除く。

【解決手段】　分野ごとに分類済みの文書を格納する分類済み文書集合格納装置２１と、この分類済み文書集合格納装置２１から取得された文書に含まれる単語の出現頻度の情報を含む分野別のテーブルを作成する分野テーブル作成部３１と、この分野テーブル作成部３１により作成された分野別のテーブルから得られる所定の語における各分野での出現頻度に基づいて、テーブルから分野別に不要語を除去する不要語決定・除去部３２と、この不要語決定・除去部３２により不要語が除去されたテーブルを格納する分類用カタログ格納装置２２と、分類される分類対象文書を格納する分類対象文書格納装置２３と、この分類対象文書格納装置２３に格納された分類対象文書に対し、分類用カタログ格納装置２２に格納されたテーブルを用いて分類処理を行なう文書分類処理部３３とを含む。

【選択図】　　図１

認定・付加情報

特許出願の番号	特願２００２－３０１５３９
受付番号	５０２０１５５４９６５
書類名	特許願
担当官	土井　恵子　　４２６４
作成日	平成１４年１０月１７日

＜認定情報・付加情報＞

【特許出願人】
- 【識別番号】 390009531
- 【住所又は居所】 アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
- 【氏名又は名称】 インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】
- 【識別番号】 100086243
- 【住所又は居所】 神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内
- 【氏名又は名称】 坂口　博

【代理人】
- 【識別番号】 100091568
- 【住所又は居所】 神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内
- 【氏名又は名称】 市位　嘉宏

【代理人】
- 【識別番号】 100108501
- 【住所又は居所】 神奈川県大和市下鶴間１６２３番１４　日本アイ・ビー・エム株式会社　知的所有権
- 【氏名又は名称】 上野　剛史

【復代理人】 申請人
- 【識別番号】 100104880
- 【住所又は居所】 東京都港区赤坂５－４－１１　山口建設第２ビル６Ｆ　セリオ国際特許事務所
- 【氏名又は名称】 古部　次郎

次頁無

出　願　人　履　歴　情　報

識別番号　　〔３９０００９５３１〕

１．変更年月日　　２００２年　６月　３日
［変更理由］　住所変更
住　所　　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
氏　名　　インターナショナル・ビジネス・マシーンズ・コーポレーション